PB83-129924

Information Flow in the Brain
Computer Requirements
(A Tutorial)

(U.S.) Health Effects Research Lab.
Research Triangle Park, NC

1982

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing) MS-82-117

1. REPORT NO. EPA-600/J-82-253	2 JOURNAL ARTICLE	3. PB83-129924
4. TITLE AND SUBTITLE Information flow in the brain: Computer requirements (A tutorial)		5. REPORT DATE 1982 6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) V.A. Benignus and K.E. Muller		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS NTD, US EPA, RTP, NC 27711 Department of Psychology, UNC-CH, NC 27514		10. PROGRAM ELEMENT NO. 11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Office of Research and Development Health Effects Research Laboratory US Environmental Protection Agency Research Triangle Park, NC 27711		13. TYPE OF REPORT AND PERIOD COVERED 14. SPONSORING AGENCY CODE EPA-600/11

15. SUPPLEMENTARY NOTES

Published in: Behavior Research Methods and Instrumentation, 14(2):294-299, 1982

16. ABSTRACT

This paper covers mathematical, statistical, and experimental design considerations needed to design a computer system for EEG data acquisition, spectrum, and cross-spectrum analyses. A tutorial approach is used in which no formal proofs are offered. Conclusions are stated simply and in summary fashion, without justifying arguments. References to formal works are given. Conclusions about system design are discussed. Frequently encountered design issues are mentioned with various alternatives.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT Release to Public	19. SECURITY CLASS (This Report) Unclassified	21. NO. OF PAGES 9
	20. SECURITY CLASS (This page) Unclassified	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

EPA-600/J-82-253
JOURNAL ARTICLE

Behavior Research Methods & Instrumentation
1982, Vol. 14 (2), 294-299

Information flow in the brain: Computer requirements (A tutorial)

VERNON A. BENIGNUS
Neurotoxicology Division, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, and Department of Psychology, University of North Carolina, Chapel Hill, North Carolina 27514

and

KEITH E. MULLER
Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27514

This paper covers mathematical, statistical, and experimental design considerations needed to design a computer system for EEG data acquisition, spectrum, and cross-spectrum analyses. A tutorial approach is used in which no formal proofs are offered. Conclusions are stated simply and in summary fashion, without justifying arguments. References to formal works are given. Conclusions about system design are discussed. Frequently encountered design issues are mentioned with various alternatives.

Measures of information flow in the brain can be derived from the electroencephalogram (EEG). Computers for measuring and analyzing EEG can be specified by consideration of requirements imposed by (1) the mathematics of estimation, (2) inferential statistics, and (3) data acquisition. In this paper, the methods of analyses are described and then the computer requirements are deduced.

The EEG is a continuously varying voltage measured from the brain. For digital analysis, this voltage wave is converted to number (digits) at each of many equally spaced points in time. This is accomplished by an analog-to-digital (A/D) converter. The result is a vector of digits for each time period (epoch) of interest in the EEG. The A/D converter must sample the continuous wave at a rate of at least twice per cycle of the highest frequency present in the wave.

EEG SPECTRA

The following is an informal presentation of spectrum analysis methods. Further reading is recommended in Bendat and Piersol (1971) and Jenkins and Watts (1968).

One-Channel Data

Assumptions. A useful model for the EEG is provided by treating it as a Gaussian (normally distributed) stochastic process. In such a process, no periodic waves are present, only waves that randomly vary about some center frequency, amplitude, and phase. The model has great appeal, since Gaussian estimation theory applies.

Send requests for reprints to Vernon A. Benignus, Environmental Protection Agency/MD-58, Research Triangle Park, North Carolina 27711.

It is convenient to define a statistical population of all EEG vectors collectable under a fixed set of relevant conditions. Relevant conditions include all variables that affect the EEG, such as external stimuli and internal states. Thus a separate population is defined for each set of relevant conditions. The objectives of EEG research may be taken to be to estimate the parameters of the populations of vectors, test differences among them, and associate the differences with relevant conditions.

Generating model waves. A wave, a vector of voltages at a series of points in time, V, can be thought of as a sum of sine and cosine waves of a series of F frequencies. An expression for the voltage at any point in time is:

$$V = V_t = \sum_{f}^{F} [a_f \cos(2\pi ft) + b_f \sin(2\pi ft)]. \qquad (1)$$

The terms inside the brackets show that at each frequency, f, a pair of waves, a cosine and a sine, are added together in proportions a_f and b_f. This produces a wave that is one of the frequency components of V. Figure 1 shows one such resultant wave, V, which is the sum of an equal proportion of cosine and sine waves of the same frequency. At Time 0, the start of the vector, a cosine wave is maximum, a sine wave is zero, and the wave V has a voltage of .5 of its maximum. By varying a_f and b_f, the relative proportion of cosine and sine waves, V can essentially be moved back and forth in time without changing its frequency or wave shape. The temporal position of a wave with respect to $t = 0$ is related to its phase. Any wave from an EEG epoch can be produced by adding a number of cosine and sine waves in different proportions and frequencies.

Another method for producing waves is to sample from a Gaussian random-number generator and plot the results as a function of time. This process is called "random" noise. Application of a filter can restrict

1982 Psychonomic Society, Inc.

0005-7878/82/020299-06$00.85/0

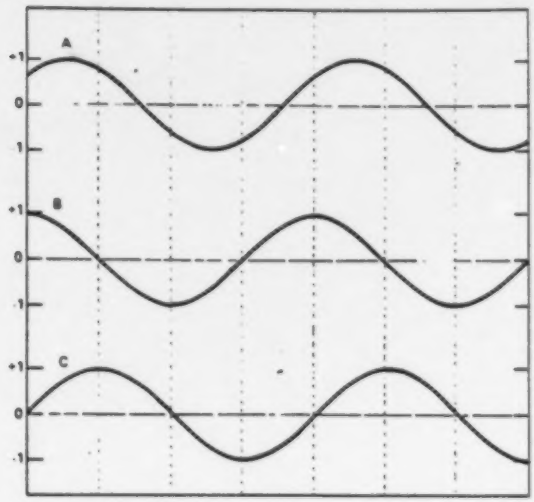

Figure 1. Waves B and C are cosine and sine waves, respectively. Their equally weighted sum is shown in A. By varying the proportion of a cosine and a sine wave at the same frequency, the phase of their sum can be varied.

the rates of oscillation. A band-pass-filtered random noise becomes a narrow-band noise. Figure 2 illustrates a purely random noise and a narrow-band random noise.

Spectra. The mean and standard deviation of an EEG vector are also called its "offset" and "amplitude." The sample vector can be "decomposed" into its various frequency components by using a form of Fourier analysis. To do this, the EEG vector is used as the dependent variable and a cosine and a sine wave at one particular frequency are used as predictors in a multiple regression. This least squares procedure estimates a_f and b_f, the amounts of cosine and sine waves of the particular frequency in V, the EEG vector. Performing a regression analysis at each frequency produces a series of pairs of regression coefficients that are estimates of the set of a_f and b_f. The a_f series is called the "cosine spectrum," and the b_f series is called the "sine spectrum."

Since it can be shown that cosine and sine waves are uncorrelated, the total activity of the wave at each frequency can be computed by

$$A_f = \sqrt{a_f^2 + b_f^2} \qquad (2)$$

or

$$P_f = a_f^2 + b_f^2. \qquad (3)$$

A_f defines the "amplitude spectrum," and P_f defines the "power spectrum." The usual procedure is to compute the $\log(P_f)$, which is in units of decibels and can be assumed to approximately follow a Gaussian distribution.

Not all frequencies can be used as predictors for a finite-length EEG vector. The lowest frequency wave that can complete a full cycle in T sec is $f_1 = 1/T$. The next highest frequency able to complete a whole number of cycles is $f_2 = 2/T$. It can be shown that sinusoids at f_1 and f_2 are uncorrelated, whereas sinusoids spaced more closely than 1/T are correlated. The highest frequency in the analysis can be no more than half the A/D conversion rate. If, for example, an EEG vector were converted at a rate of 100 samples/sec and its length were 2 sec, then the lowest frequency available in the spectrum would be .5 Hz, the resolution (spacing of frequencies) would be .5 Hz, and the highest frequency resolved would be 50 Hz.

Stability of estimates. The estimate, P_f, is extremely unstable. In order to stabilize it, averaging is required. One solution is to compute the spectrum of each of several EEG vectors and then average across the spectra to produce more stable estimates. Another solution is to average across the several adjacent values of P_f within one spectrum. While the latter method reduces the number of frequencies in the spectrum, it is frequently the case that only a reduced set is needed anyway (e.g., Benignus, Otto, Muller, & Seiple, 1981). Both forms of averaging can be done in the same estimate.

Two-Channel Data

Assumptions. If data are collected from two sites, A and B, on the same subject at the same time, it is reasonable to ask about the relationship between the two EEG vectors. To do so, it is customary to assume that EEGs at A and B are linear functions of each other via some mechanism, a "data transmission path." Such a hypothetical path might be a set of multiple connections or even a third site that sends signals to both A and B. The data being transmitted could consist of neural impulses, graded potentials, or chemical events.

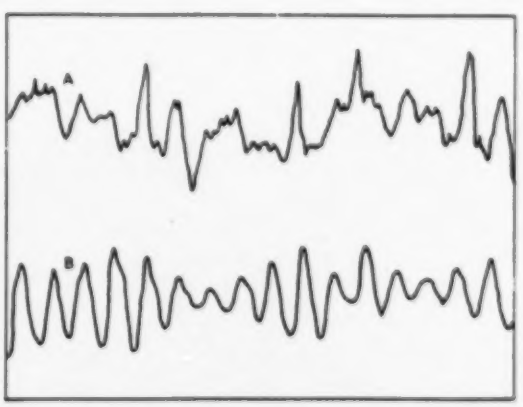

Figure 2. A is a purely random noise; B is a narrow-band noise.

Coherence. Measures of correlation between spectra of EEG vectors from A and B, for each frequency in the spectrum, allow discovery of relationships that occur only in one frequency. Simply because power in a particular frequency band from A correlates with power in the same band from B does not necessarily imply signal coupling. A more stringent requirement is that the waves in each frequency from A and B must also be synchronized (time locked).

The two Fourier coefficients of the spectrum, a_f and b_f, which are squared and added to produce P_f, provide information about phase. By computing correlations between a_{fA} and a_{fB} across trials at each frequency and another set of correlations between b_{fA} and b_{fB}, one arrives at two spectra of correlations, R_{fa} and R_{fb}. Since the two correlations were computed from uncorrelated variables, a new quantity, describing the total relationship at each frequency, may be defined as:

$$\gamma_f^2 = R_{fa}^2 + R_{fb}^2. \quad (4)$$

The set of quantities γ_f^2, the "coherence," is a squared multiple-correlation spectrum that describes the relationship between the two EEG vectors at each frequency in the spectrum. This computation uses one pair of spectra from each of several trials as data to compute the correlations.

Regression coefficients. It is also useful to compute regression coefficients that are the least squares estimates of slope for the relationship between A and B. Consistent with the computation of R_{fa} and R_{fb}, Slopes S_{fa} and S_{fb} may be combined into one slope spectrum describing the total relationship at each frequency as:

$$G_f = \sqrt{S_{fa}^2 + S_{fb}^2}. \quad (5)$$

In network analysis the quantity G_f is the "gain" of the transmission path. Since the a_f and b_f components (cosine and sine) were handled separately, however, and since least squares estimation was used, G_f represents the gain of the path for synchronized signals only.

Phase. Since phase information for A and B waves is retained by a_f and b_f, the timing relationships between the waves from A and B may be described. This spectrum, the phase spectrum ϕ_f, measured in radians, gives the lead or lag of the output vs. the input at each frequency.

Summary of Estimation

One can estimate the spectra of a population of EEG vectors. By computing γ_f^2, G_f, and ϕ_f, the relationship, gain, and timing of a signal transmission path between A and B can be estimated for each frequency.

INFERENTIAL STRATEGIES

Role of Statistical Thinking

Once the data have been collected, the emphasis shifts to ways of estimating and testing relationships with treatments. Four major issues will be discussed: (1) the role of statistical thinking, (2) design strategies, (3) statistical issues in choosing measures, and (4) appropriate analysis techniques.

The nature of EEG technology makes it easy to collect large amounts of data. It is important to consult with a statistician before data are collected. By focusing on a particular set of goals, the tasks of data reduction and data analysis are made easier.

It is important to remember a distinction between exploratory analysis and confirmatory analysis. Confirmatory analysis involves a study in which testing procedures and hypotheses to be tested are all chosen a priori, before data collection. Once data have been collected, the fixed hypotheses are tested with fixed techniques, resulting in meaningful probability values (p values). Analyses that do not fit this model are exploratory. EEG data demands both exploratory and confirmatory analysis. Proper design strategies must be used to assure correct assignment of p values for confirmatory analyses. Problems associated with controlling Types I and II error rates in experimental design are important in this context.

Design Strategy

The design solutions to be suggested are split-samples techniques. With this technique, one treats different parts of the data differently. In a replication study, one first collects data and does exploratory analysis. Based on the conclusions, one then conducts a replication study. Such an approach requires a great deal of work. It is usually more efficient to run a small pilot study, but this does not save much work. An alternate related approach is a "leapfrog design." A leapfrog design involves planned hypotheses and tests that are published. However, additional measures are collected and analyzed but are not published. The extra measures and exploratory analyses are used as pilot data to design a subsequent study. The effort necessary to collect extra measures during or after the original session is quite small. The planned study designed from the added exploratory data analysis may serve as a basis to collect pilot data for a subsequent study.

Choosing Measures

P_f, γ_f^2, G_f, and ϕ_f may be computed for a number of frequencies. Each measure at each frequency is a separate dependent variable. The naive approach would be to do a separate analysis on each. Unfortunately, this would lead to far too many Type I errors. The reasons

for this are the same as those for doing an overall F test rather than a collection of t tests. This fundamental problem demands variable-reduction schemes (i.e., ways to choose as few variables as possible).

In order to insure that important variables have not been missed, one may do exploratory analysis and then do a follow-up study with only the "significant" measures. Another way to choose variables is from previous experience.

More formal variable-reduction schemes exist, such as some form of factor analysis. Principal-components techniques are likely to be preferred over others, since (1) the original metric is often of interest and (2) the purpose is to produce new scores that are linear combinations of original variables. Factor analysis should be done on data either only from a control group or only from residuals. Here, "residuals" refers to responses from which any treatment effects have been statistically removed. If this is not done, then the presence of experimental design effects may badly distort the structure (Muller, Hosking, & Helms, 1979).

Other exploratory variable-reduction techniques include stepwise regression and stepwise discriminant analysis. Stepwise discriminant may be used to find those variables that best predict treatment effects.

A confirmatory technique may be used as an exploratory analysis as long as one does not confuse the purpose and the results. By conducting a univariate analysis of variance (ANOVA) for each different measure, one may compare the nominal p values and the observed correlations between response and predictors. Possible response candidates can then be chosen. If the subsequent study (independent of those data) uses only the chosen variables, then a valid leapfrog design has been used.

In a multivariate situation, a multivariate ANOVA (MANOVA) may be used to test different candidate collections of responses. In an exploratory study, anything that may gain information is appropriate. However, p values in such a context cannot be reported as confirmatory p values. Rao (1973) suggested a different exploratory method in the multivariate context for evaluating redundancy of candidate response variables.

A priori reduction rather than the above post hoc methods may also be useful. One important method is to define broad frequency bands so as to define fewer measures. One may also choose just a few frequencies or fewer electrodes. In most cases, such approaches will only be satisfying in the context of a priori pilot data.

Inferential Techniques

Once data have been collected and derived measures computed, a confirmatory study requires formal hypothesis tests to decide if any differences are significant. For most cases, MANOVA is the preferred approach, due to the likelihood of violation of the assumptions of univariate repeated-measures ANOVA. Violations of assumptions usually yield anticonservative results. (Corrections for repeated-measures ANOVA results are reviewed by Kesselman and Rogan, 1980.)

Distributional problems may arise with ϕ_f and γ_f^2. The distributions of these variables especially must be checked. Transformations will usually lead to less useful data, except for the log transform of P_f, which is a decibel scale.

Multiple t tests in lieu of an ANOVA have nearly been eliminated from the psychological literature. The same problem arises with multiple ANOVAs. In order to provide proper error rate protection, one should either do MANOVA or provide Bonferroni correction for the nominal p values (Abt, 1981). The Bonferroni approach is least conservative when the measures overlap the least. Useful references to statistical material are found in Timm (1975).

COMPUTER REQUIREMENTS

No effort will be made to consider specific hardware or software details. Attention will be given to performance requirements based on the computations needed for estimation and testing. Table 1 gives a classification scheme for computers used in EEG analysis. As a rule, a

Table 1
A Classification Scheme for Computers

Class	Processor	Peripheral Devices	Operating System	Languages
Micro	Often one chip on one board	Cassette, floppy disk, terminals, slow printers	?	Assembler, interpreter for BASIC, PASCAL
Mini	Faster CPU, faster interfaces for peripherals	Some form of almost any device	Batch OS, system utilities	BASIC, assembler, interpreter and compiler for FORTRAN
Midi	Even faster CPU and interface, may have subsidiary micros, virtual memory	Can run almost any small collection, smart peripherals	Maybe timesharing, excellent utilities, choice of systems	Optimizing compilers, powerful assembler, software statistical packages
Maxi	Multiprocessor, typically virtual memory	A confederation of very smart collections of devices	Batch and timesharing—an environment, not a "computer"	Limited only by mainframe manufacturer

type of computer may be able to do things its near neighbors can, but it typically will not be able to do some things its far neighbors can. Computer requirements will be discussed within the framework shown in Figure 3. The process blocks are arbitrary, but processing is sequential, so the ordering of steps is approximately fixed.

Before A/D Conversion

Process control. The choices for process control range from a separate micro, dedicated to control, through a midi that performs both control and data acquisition (A/D conversion). Techniques of process control are clearly beyond the goals of this paper. Consideration of device time contention and speed, discussed in the context of other functions, also apply here.

Analog preprocessing. This important step involves considerations such as impedance matching, prefiltering, noise, and amplification. Problems of this class are sometimes undetected in a working system. Design and maintenance must involve personnel who are familiar with (1) data, (2) the purpose of the research, (3) engineering, and (4) computation.

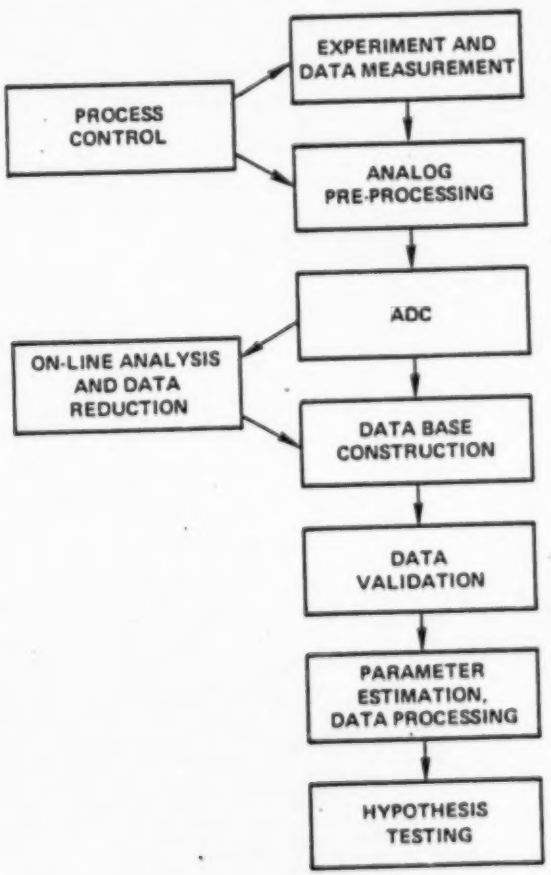

Figure 3. A diagram of data processing steps involved in EEG experiments on spectra and cross-spectra.

A/D Conversion

The A/D conversion rate must be twice the highest frequency present in the data. For EEG, A/D conversion rates from 125 Hz up are useful, but for some other purposes, such as sensory evoked potentials, much higher rates, as high as 10 kHz, are necessary.

To convert several EEG channels "simultaneously," a multiplexer is often used. For a K-channel system the multiplexer rapidly switches (scans) the A/D converter through all K channels and the A/D converter converts one value from each channel in turn. Thus, a set of K samples (a frame) is collected, each of which represents the first point in K vectors. After waiting for a time determined by the sampling rate, another frame is collected, forming the second point in K vectors. The rate at which the multiplexer moves from channel to channel within a frame is the scan rate. The sampling rate is the frame rate.

The scan rate must be high compared with the frame rate unless a delay can be tolerated between EEG channels. If analysis of paths between channels is part of a study, then delays could give false information about phase shift and reduce γ_{ij}^2, especially at high frequencies. Nonzero delays between channels should at least be constant. One solution to the problem of slow scan rates is the use of an analog sample-and-hold device at the input of each multiplexer channel. In such a system, all K sample-and-hold devices take a sample simultaneously. The scan rate must then only be high enough to finish before the next frame.

Microprocessors can serve as data receiving machines (hosts) in EEG work only if special direct memory access (DMA) hardware is available; otherwise, data transfer (baud) rate is too low. Baud rates are usually high enough in non-DMA hosts of minicomputer size or larger. When a maxi or midi is used, however, a timesharing system frequently becomes a problem due to contention for cycle time. Many midi manufacturers offer A/D converter units with a dedicated micro and memory to overcome contention problems. In general, real-time data collection on a maxi is difficult or impossible. Users should be cautioned that even though the peripheral A/D conversion system has sufficient speed, this does not mean that a particular host and operating system can handle the data rate.

Mass Data Storage

Raw A/D converter data may be stored "as is" or reduced on line. Storage of raw data requires fast mass storage devices. Additional complexity in programming and operation for off-line analysis occurs if a mass storage device cannot contain the data from all subjects. When a multivariate and/or a multifactor design is used, data from a large number of subjects must be collected

to gain sufficient power. To compute stable spectrum estimates, many trials must be collected for averaging. These requirements demand large devices. With today's technology, a data base may be too large for most disk packs.

An alternative to keeping raw data is to do on-line data reduction before storage. Reduction may include rejecting artifact-contaminated data and computing spectra to be averaged across trials. Also two or more adjacent A/D converter samples may be averaged. If J adjacent samples are averaged, the EEG vector size is reduced by a factor of J, as is the storage needed. The analog data must still be sampled at a rate high enough to avoid certain errors (Bendat & Piersol, 1971, chap. 7).

The more on-line reduction that occurs, the less off-line analysis flexibility exists due to inappropriate data loss. On-line reduction is least appropriate during pilot work and may preclude a leapfrog strategy. If speed, size, and cost become limiting factors, on-line reduction becomes mandatory.

Data Validation

Validation of data stored in the data base is an often neglected but important step. Validation demands some mix of (1) hand editing, (2) software, and (3) controlled experimental conditions. These elements define a three-dimensional performance envelope. Artifact detection can be mostly automated, given sufficient knowledge of the paradigm. Plots of raw data allow the experimenter to check the functioning of the A/D converter, detect additional artifacts, and so on. Validation must be done trial by trial. It may be done on either the acquisition computer or the data analysis computer. Ease of use and availability of software dominate this decision.

Parameter Estimation

Spectrum calculation. A vast literature exists on methods for spectrum calculations. Ideally, this should be done on a large computer using public, packaged software, although much software is also available on mini and midi systems. BMDP81 and SAS both have good programs and features. Buying and using packaged software is almost always effectively cheaper. Writing one's own software increases the possibility of undetected errors.

Spectrum averaging is simple on almost any computer. Data analysis systems such as BMDP, SAS, and SPSS provide convenient methods for averaging over trials to form stable estimates.

Filtering. Removing ranges of frequencies from V before spectrum calculation can improve estimation. Strong signals in narrow frequency ranges can contaminate estimates of weaker signals. Many techniques for digital filtering are available (e.g., in BMDP81).

Inferential Statistics

This step should be reserved for general-purpose statistical packages on at least a midi, and preferably a maxi. The steps involve (1) data validation, (2) descriptive statistics, (3) hypothesis tests, and, perhaps, (4) data displays, such as plots.

Data management occurs throughout the steps of this process. This includes creating files, storing files, creating new variables and file formats, and documenting analysis. Statistical packages help do this and also provide easy access to analysis software.

Data validation at this point is demanded, since many errors in earlier steps are easy to find here. Range checks, means, variances, histograms, correlations, and scattergrams are easy to compute and helpful.

Much of the hypothesis testing should be done with general linear models programs. These include MANOVA, regression, discriminant analysis, and canonical correlation. Factor analysis programs may be needed. Again, the larger the computer and more powerful the package, the easier analysis is.

CONCLUSIONS

A complete design of a computing system for EEG laboratory work should not be approached piecemeal. It is possible to design data acquisition and storage strategies that make subsequent analyses quite difficult. This does not imply the use of one machine to do all jobs. Although possible, it is unlikely to be economical. A mixed system involving microcomputer-controlled peripherals, a mini or midi host, and a data analysis maxi may be optimal.

REFERENCES

Abt, K. Problems of reported significance testing. *Controlled Clinical Trials*, 1981, 1, 377-381.

Bendat, J. S., & Piersol, A. G. *Random data: Analysis and measurement procedures*. New York: Wiley, 1971.

Benignus, V. A., Otto, D. A., Muller, K. E., & Seiple, K. J. Effects of age and body lead burden on CNS function in young children. II. EEG spectra. *Electroencephalography and Clinical Neurophysiology*, 1981, 52, 240-248.

Jenkins, G. M., & Watts, D. G. *Spectral analysis and its applications*. San Francisco: Holden-Day, 1968.

Keselman, H. J., & Rogan, J. C. Repeated measures F-tests and psychophysiological research: Controlling the number of false positives. *Psychophysiology*, 1980, 17, 499-503.

Muller, K. E., Hoeking, J. D., & Helms, R. W. Using LINMOD to adjust for treatment effects when analyzing the covariance matrix. *Proceedings of the Statistical Computing Section*, ASA, 1979, 136-139.

Rao, C. R. *Linear statistical inference and its applications*. New York: Wiley, 1973.

Timm, N. H. *Multivariate analysis with applications in education and psychology*. Belmont, Calif: Wadsworth, 1975.